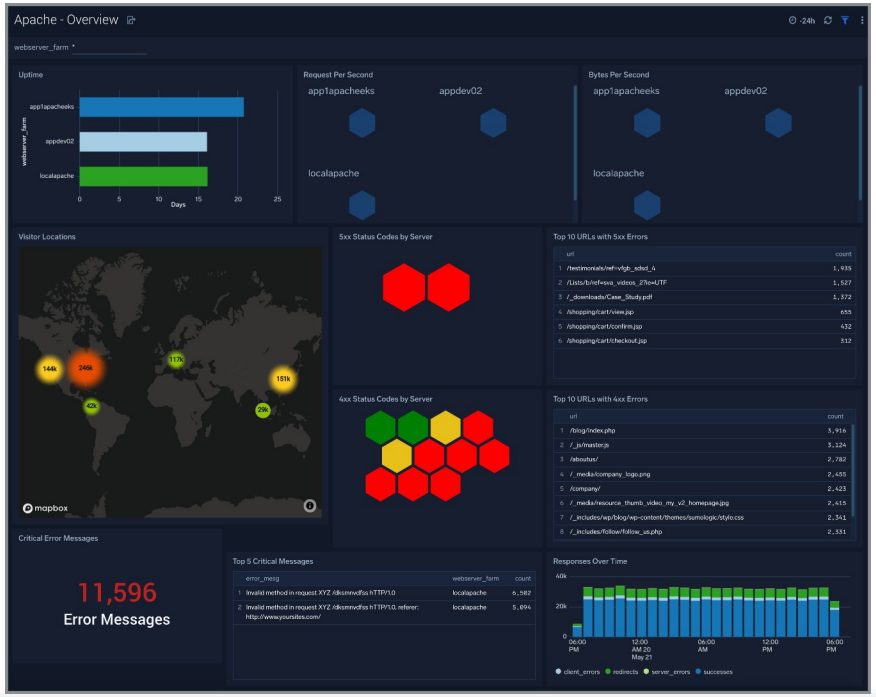
Administration système

Travaux dirigés

Extraction et traitement de données d’un fichier de log Apache

Outils de reporting Apache

- Le contexte
  - Il existe des outils de reporting payants mais votre boss préfère vous demander d'en fabriquer un moins conséquent que celui-ci mais...
    - Gratuit et
    - 'Adapté au contexte de votre entreprise'



https://www.sumologic.com/application/apache/

Outils de reporting Apache

- Le besoin
  - Votre boss voudrait un script Python qui lui permette de connaître les connexions au serveur Apache avec les données suivantes :
    - l’adresse IP
    - le status de la réponse
    - le nombre d’octets envoyés
    - Tout cela regroupé par @IP

Résultat

- Quel est le résultat attendu ?

```
# affichage de chaque ligne du fichier
{'remote_host': '127.0.0.1', 'status': '200', 'bytes_sent': '2326'}
{'remote_host': '192.168.0.1', 'status': '200', 'bytes_sent': '4567'}
{'remote_host': '111.222.333.123', 'status': '200', 'bytes_sent': '198'}
{'remote_host': '111.222.333.123', 'status': '200', 'bytes_sent': '28083'}
{'remote_host': '111.222.333.123', 'status': '200', 'bytes_sent': '9332'}
{'remote_host': '111.222.333.123', 'status': '200', 'bytes_sent': '207'}
# traitement pour regroupement par @IP
{'127.0.0.1': [2326], '192.168.0.1': [4567], '111.222.333.123': [198, 28083, 9332, 207]}
```

Soit le fichier de log suivant :

acces1.log (*au format Combined Log Format*)

```
127.000.000.001 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
192.168.000.001 - Joe [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 4567 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
111.222.333.123 HOME - [01/Feb/1998:01:08:39 -0800] "GET /bannerad/ad.htm HTTP/1.0" 200 198 "http://www.referrer.com/bannerad/ba_intro.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
111.222.333.123 HOME - [01/Feb/1998:01:08:46 -0800] "GET /bannerad/ad.htm HTTP/1.0" 200 28083 "http://www.referrer.com/bannerad/ba_intro.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
111.222.333.123 AWAY - [01/Feb/1998:01:08:53 -0800] "GET /bannerad/ad7.gif HTTP/1.0" 200 9332 "http://www.referrer.com/bannerad/ba_ad.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
111.222.333.123 AWAY - [01/Feb/1998:01:09:14 -0800] "GET /bannerad/click.htm HTTP/1.0" 200 207 "http://www.referrer.com/bannerad/menu.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
```

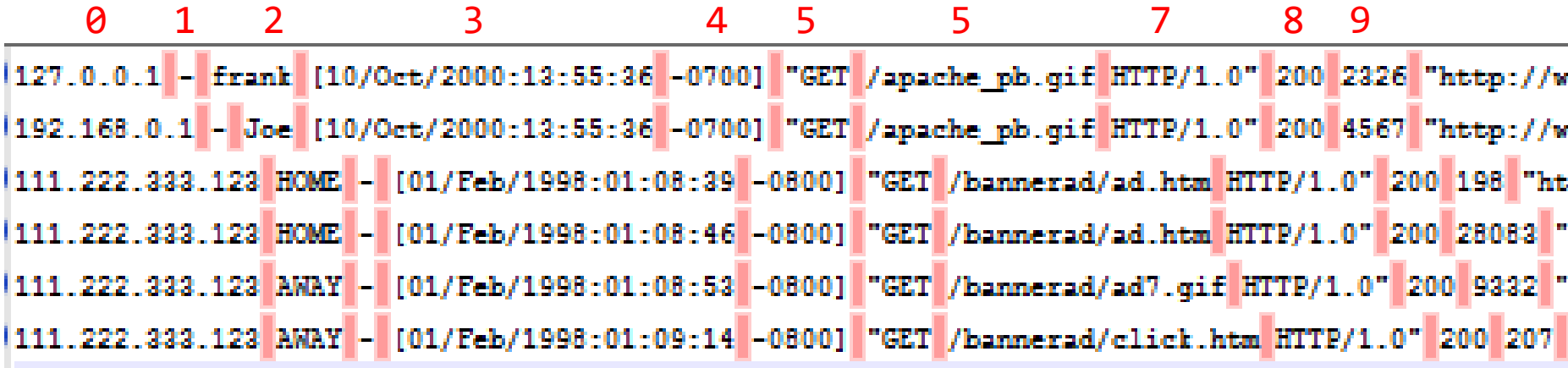
```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
192.168.0.1 - Joe [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 4567 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
111.222.333.123 HOME - [01/Feb/1998:01:08:39 -0800] "GET /bannerad/ad.htm HTTP/1.0" 200 198 "http://www.referrer.com/bannerad/ba_intro.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
111.222.333.123 HOME - [01/Feb/1998:01:08:46 -0800] "GET /bannerad/ad.htm HTTP/1.0" 200 28083 "http://www.referrer.com/bannerad/ba_intro.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
111.222.333.123 AWAY - [01/Feb/1998:01:08:53 -0800] "GET /bannerad/ad7.gif HTTP/1.0" 200 9332 "http://www.referrer.com/bannerad/ba_ad.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
111.222.333.123 AWAY - [01/Feb/1998:01:09:14 -0800] "GET /bannerad/click.htm HTTP/1.0" 200 207 "http://www.referrer.com/bannerad/menu.htm" "Mozilla/4.01 (Macintosh; I; PPC)"
```

```
0   1   2          3                    4   5          5          7        8   9
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://w
192.168.0.1 - Joe [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 4567 "http://w
111.222.333.123 HOME - [01/Feb/1998:01:08:39 -0800] "GET /bannerad/ad.htm HTTP/1.0" 200 198 "ht
111.222.333.123 HOME - [01/Feb/1998:01:08:46 -0800] "GET /bannerad/ad.htm HTTP/1.0" 200 28083 "
111.222.333.123 AWAY - [01/Feb/1998:01:08:53 -0800] "GET /bannerad/ad7.gif HTTP/1.0" 200 9332 "
111.222.333.123 AWAY - [01/Feb/1998:01:09:14 -0800] "GET /bannerad/click.htm HTTP/1.0" 200 207
```

https://httpd.apache.org/docs/2.4/logs.html#accesslog

Données significatives

- Les données qui nous intéressent sont :
  - 0 : l'adresse IP
  - 8 : le status de la réponse
  - 9 : le nombre d'octets envoyés
  - L'ensemble des octets regroupés par @IP



Processus

- Quel est le processus de création d'un script Python qui va donner ce résultat ?
- Le script doit :
  - être appelé depuis la ligne de commande
  - Vérifier qu'au moins un argument a été passé
  - Prendre en argument le nom du fichier de log (sys.argv)
  - Vérifier que le fichier existe (exception)
  - Fonction 1 : parcourir toutes les lignes du fichier (boucle for)
    - Fonction 2 :
      - décomposer la ligne (split)
      - Créer un dictionnaire avec trois clés
      - Afficher ce dictionnaire
    - *Faire le regroupement par adresse IP (méthode setdefault)*